基于增量队列的在全置信度下的关联挖掘①

刘 炜

(上海交通大学 管理科学与工程, 上海 200030)

摘 要: 关联挖掘是一种重要的数据分析方法, 提出了一种在全置信度下的增量队列关联挖掘算法模型, 在传统的 FP-Growth 及 PF-Tree 算法的关联挖掘中使用了全置信度规则, 算法的适应性得到提升, 由此提出 FP4W-Growth 算法并运用到对文本数据的关联计算以及对增量式的数据进行关联性挖掘的研究中, 通过实验验证了此算法及模型的可行性与优化性, 为在庞大的文本数据中发现隐藏着的先前未知的并潜在有用的新信息和新模式, 提供了科学的决策方法.

关键词: 海量文本; 增量队列; 全置信度; 关联挖掘

Association Mining on Massive Text under Full Confidence Based on Incremental Queue

LIU Wei

(Management Science and Engineering, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: Association mining is an important data analysis method, this article proposes an incremental queue association mining algorithm model under full confidence,using the full confidence rules in the traditional FP-Growth and PF-Tree association mining algorithm can improve the algorithm adaptability. Thus, the article proposes FP4W-Growth algorithm, and applies this algotithm to the association calculation of text data and association mining of incremental data. Then this paper conducted verification experiment. The experimental results show the feasibility of this algorithm and model. The article provides a scientific approach to finding hidden but useful information and patterns from large amount of text data.

Key words: large amounts of text; incremental queue; full confidence; association mining

在互联网和计算机快速发展的今天, 虚拟世界中的数据呈现爆炸式增长, 文本信息也伴随着信息技术的发展形成了大量的数据累积, 随之而来的便是如何从这些大量的信息中获得符合需求的或潜在有用的数据和知识. 面对处理海量文本信息时乏而无效的手段, 数据挖掘技术的出现与发展, 不但从理论上提出了海量数据的分析方法, 更从技术上将理论用于实践, 产生了一大批数据挖掘、关联分析模型, 为更好的从数据中获得知识提供了创造性解答.

关联挖掘的算法优化与研究已经成为数据挖掘领域中一门重要的研究课题. Rakesh Agrawal 等人在首先提出了 Apriori 算法, 旨在知识发现领域运用关联规则挖掘. 再此基础上, 后来人进行了一系列针对性的优化和改进, 提出增量关联规则研究, 在增量式的文本信息中运用关联规则挖掘, 从信息中发现频繁片断的算法[1]. Tung 等人[2]提出一种先挖掘信息内关联规则, 然后挖掘增量信息间关联规则的 FITI(First Intra Then Inter)算法[3]. 但由于这种算法基于 Apriori 算法, 其性能瓶颈也体现在会生成大量后选项集和需要频繁扫描整个数据库. Lu[4,5]等人提出了关联规则挖掘算法 E-Apriori 和 EH-Apriori 算法. 以上都是以 Apriori 算法为基础, 其中 EH-Apriori 算法利用哈希算法产生了频繁 1 项集, 然后利用频繁 1 项集生成备选 2 项集, 去掉不合适的候选 2 项集, 所以它的效率比 E-Apriori 算法高. 为了获得更高的效率 Han[6]提出了一种基于 FP-Tree 的不产生候选的 FP-Growth 算法, 它可以生成

① 收稿时间:2014-11-26;收到修改稿时间:2015-01-19

一个压缩的 FP-Tree, 并压缩原来的数据库. 这种算法关注频繁模式(段, 子序列)增长, 不会生成多余的候选项, 效率得到了提高[7]. 范明等基于 FP-Growth 算法提出了一种不生成条件 FP-Tree 的挖掘算法, 提高了频繁模式挖掘的时空效率[8].

本增量尺度 FP-Growth 算法的思想, 利用全置信度的性质对挖掘任务进行了改进, 使算法更适合处理信息数据, 最后给出了算法性能的实验结果.

1 相关定义及描述

以下介绍一些增量队列关联挖掘算法中出现的概念, 其中涉及文本特征词集、增量词集队列、文本增量区间这样的基础概念, 以此为基础阐述了增量关联规则和全置信度, 最终引出了全置信下的空间剪枝规则, 能够有效提高搜索有效性, 进而提高算法时间效率.

1.1 文本特征词集

某类型数据为文本数据, 对于给定数量为 S 的文本数据, P 为定量文本集, 由其文本集产生的特征词按权重区间集 $Q=\{Q_1, Q_2, Q_3, \ldots\ldots Q_n\}$归纳为有序项队列.

$W=\{S_i, W_1, W_2, W_3, \ldots\ldots W_n\}$为在 Q 参照下的按区间分特征词集序列, 其中 Si 为产生本次特征词集序列的给定文本数据量, Wi 表示符合权重区间 $Q=Q_1$ 的特征词集合.

1.2 增量词集队列

对于首次给定数量为 S 的文本数据, W 为该次特征词集序列, 其后继续给定多次, 且每次数量为 S_i 的文本数据, 产生出的特征词集序列为 W_i.

则定义增量词集队列 $K=\{W_1, W_2, W_3, \ldots\ldots W_n\}$.

1.3 文本增量区间

基于每次给定数量为 S_i 的文本数据, 产生的本次特征词集序列 W_i 为从 $\sum_{o}^{i}(S_i)$ 到 $\sum_{o}^{i}(S_{i-1})$ 数量区间内的特征词集序列.

1.4 增量关联规则[3]

根据上述定义, 增量词集队列 K 中, 每个特征词集序列 W 对应一个增量区间 S. 我们以某段区间长为 S 的词集序列为原点, 如果在区间内, 词集序列 W 中有特征词出现, 则将该特征词标注在 W 中, 并记录为 R_iw_j.

多时间序列跨事务关联规则的支持度 Sp, 可信度 Cf 定义为:

$$Sp=\frac{C_{AB}}{n}\ ,\ Cf=\frac{C_{AB}}{C_A}$$

其中 A 出现的次数为 C_A, $A\cup B$ 出现的次数为 C_{AB}, n 为特征词集序列数.

文本信息的增量关联规则满足下列条件:

(1) $A\subset K, B\subset K, X\cap Y=\Phi$

(2) $\exists\ R_iw_j\in A, 1\le j\le n$

(3) $\exists\ R_iw_j\in A,\ 1\le j\le n, ((i=j)\wedge(1\le i<Si))\vee((i\ne j)\wedge(0\le i<Si))$

(4) $\exists\ R_iw_j\in B, 1\le j\le n, max(j)<i\le Si$

同时, 该增量关联的规则蕴涵式形式为 $A\Rightarrow B$.

1.5 全置信度

以往的支持度-置信度模型, 在海量文本信息增量叠加模式下, 尤其是在支持度相对较低时, 对无用规则的排除显得无可奈何. 同时, 也无法很好的处理文本信息数据中频繁出现的负相关等现象. 挖掘质量是关联挖掘的灵魂, 为了使质量更加可靠, 我们采取全置信度的方式来重构该模型.

我们定义全置信度用 Acf 表示, 对于给定项集 $D=A\ B, D$ 的全置信度为:

$$Acf(D)=\frac{Sup(D)}{max\_item\_sup(D)}=\frac{Sup(A\cup B)}{max\_item\_sup(A\cup B)}=\frac{Sup(A\cup B)}{max\{Sup(R_i)\mid\forall R_i\in(A\cup B)\}}$$

其中, $max\{sup(R_i)\mid\forall R_i\in(A\cup B))\}$是$A\cup B$中所有项的最大(单个)项支持度.

全置信度有两点性质:

① 零不变性. 即它的值不受空数据的影响.

② 向下封闭性. 如果一个模式全置信, 它的子模式也全置信. 如果一个模式非全置信的, 那么该模式增长后也达不到最小全置信度阈值[7].

1.6 空间剪枝规则

一个频繁模式 F 是全置信的, 必须满足两个条件:

(1) $Acf(F)min\_$

(2) $Sup(F)min\_$

由此引出空间剪枝规则: 项集 $\theta=K_1, K_2, \ldots\ldots K_n$, 在被 θ 约束子树中[8], 对于全置信模式 F 有

$$Sup(F)\le\frac{Sup(\theta)}{min\_\theta}$$

证明: θF 是全置信的,

则 $all\_conf(\theta F) \geq min\_\theta$

$$\Rightarrow \frac{Sup(\theta F)}{max\_item\_sup(\theta F)} \geq min\_\theta$$

$$\Rightarrow max\_item\_sup(\theta F) \leq \frac{Sup(\theta F)}{min\_\theta}$$

$$\because \ |Sup(\theta)| \geq |Sup(\theta F)|$$

$$\Rightarrow max\_item\_sup(\theta F) \leq \frac{Sup(\theta)}{min\_\theta}$$

又$\because max_item_sup(F) \leq max_item_sup(\theta F)$

$$\Rightarrow max\_item\_sup(F) \leq \frac{Sup(\theta)}{min\_\theta}$$

2 增量队列关联挖掘算法

FP4W-Growth 算法, 主要基于范明的不生成条件子树的 FP-Tree[8]的思想, 对 FP-Tree 的构造过程进行了改进, 以适用于文本增量队列的关联挖掘. 同时, 为降低开销, 采用了闭频繁集构建 FP-Tree. 此外利用引理 1 改进空间剪枝规则, 使用全置信模型构造出高度可用的关联规则.

算法具体思想如下:

构造 FP-Tree:

第一步: 找出满足最小支持度阈值的文本数据, 利用哈希算法产生频繁 1 项集.

第二步: 基于分而治之框架. 对每个参考时间基准点项 R_iw_0, 执行如下操作:

(1)当 R_iw_0 出现时, 增量队列中的频繁 1-项集 $Fre\_Item\_S_1$

(2) 将 $Fre\_Item\_S_1$ 按照 $(R_{i+1}w_0,\ldots R_nw_0, R_1w_1,\ldots R_nw_0,\ldots, R_1w_{m-1},\ldots, R_nw_{m-1})$的次序排列.

(3)读入增量词集队列 K, 每个增量队列作为一个事务, 使用项合并剪枝法[7]得到事务内的闭频繁项集 Clo_FiSet.

(4)将 Clo_FiSet 中的项按支持度降序排序.

(5)对每个增量队列创建一个分支, 构建 FP-Tree.

关联挖掘算法改进:

挖掘算法主过程参看文献[8]. 在更长全置信模式递归挖掘过程中, 设 FP-Tree 中已经存在一个长度为 k 的全置信模式 $\{F_1,\ldots F_k\}$, 如果 $Sup(F_k) \leq Sup\{ F_1,\ldots F_k \}/min\_\theta$ 并且 F_k 全置信度 $all\_conf\ min\_\theta$,就产生更长的全置信频繁模式. 最后, 在$\{Sup,Conf,Acf\}$的度量框架下, 生成关联规则, 形如: $A \Rightarrow B[support,confidence,all\_confidence]$.

3 系统仿真

根据某研究所文本信息的 11 万条数据, 总计约有 3000 多个有效词, 200 多个有用词. 在实验室环境下, 对 FP4W-Growth、FP-Growth 和 E-Apriori 三种算法的性能进行对比.

通过计算特征词, 11 万条数据为增量队列总长, 增量尺度为 1 万条数据. 我们设定系统的最低全置信度为 80%, 全置信频繁项最大长度为 8, 最低置信度为 80%, 最低支持度为 5%.

仿真结果为 FP4W-Growth、FP-Growth、E-Apriori 和 EH-Apriori 两种算法产生的关联规则量对比, 以及 FP4W-Growth、FP-Growth、E-Apriori 和 EH-Apriori 四种算法的在空间和时间的性能对比, 如下图所示.

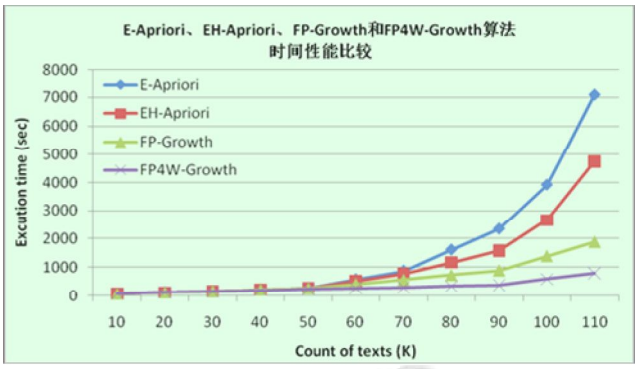

图 1 E-Apriori、EH-Apriori、FP-Growth 和 FP4W-Growth 算法时间性能比较

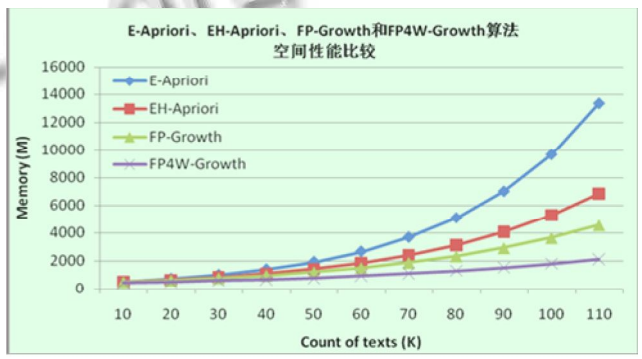

图 2 E-Apriori、EH-Apriori、FP-Growth 和 FP4W-Growth 算法空间性能比较

从图 1、图 2 中可以明显看出 E-Apriori 和 FP-Growth 算法, 信息量 5 万条以下时, 效率相差不大, 且都比较低. 随着数据增长, E-Apriori 算法和 FP-Growth 算法占用内存过多以助于超出了机器物理

内存的大小，因而必须使用虚拟内存，但大量运行时间会消耗在内存换页上. 而 EH-Apriori 算法采用哈希算法，数据量增加导致内存占用量上升，虽然占用量接近物理内存上线，但明显低于 E-Apriori 算法，因此虚拟内存使用量减少，时间效率会较高. 而 FP4W-Growth 算法因采用多项优化技术，内存占用少，内存占用随时间增长趋势平缓. 由此可见，FP4W-Growth 算法在时间/空间效率上要优于 EH-Apriori 算法与 FP-Growth 算法.

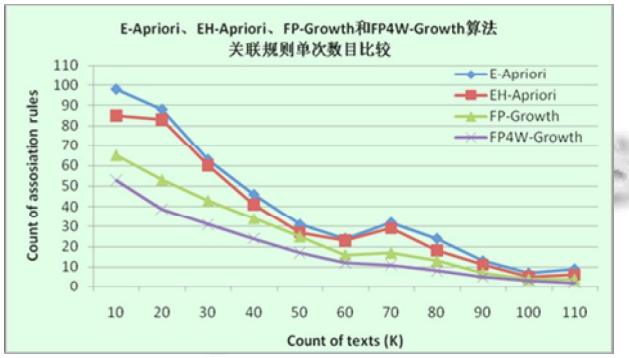

图 3 E-Apriori、EH-Apriori、FP-Growth 和 FP4W-Growth 算法单次生成关联规则数目比较

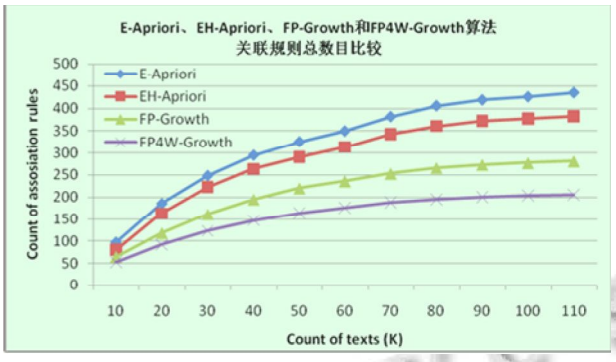

图 4 E-Apriori、EH-Apriori、FP-Growth 和 FP4W-Growth 算法生成关联规则总数目比较

从图 3、4 中可以看出，E-Apriori、EH-Apriori 和 FP-Growth 三种算法产生的关联规则数较多，其中有很多是没有研究价值的规则. 而 FP4W-Growth 算法由于采用全置信度过去掉了大量不相关的规则，提高了关联规则的质量.

4 结语

本文针对海量文本信息数据的特点，研究了增量队列的关联规则挖掘方式，用以预测与发现学科知识之间的相互运用及发展趋势，提出了基于增量队列的海量文本信息在全置信度下的 FP4W-Growth 算法. 该算法利用全置信度性质，更加有效的对不产生条件子树 FP-Tree 进行搜索、剪枝，时间和空间效率大大提高，并挖掘出兴趣度更高的规则. 在论文最后进行了一个基于增量关联规则挖掘的仿真对比及评价，证明了算法的可行性与优化性.

论文所做的工作对于海量文本信息分析中的面向分类预测的数据挖掘技术，如进行海量信息分类管理、多学科关系分析、推进不同知识的穿插应用等方面都具有一定的指导和借鉴意义.

参考文献

1 潘定,沈钧毅.态数据挖掘的相似性发现技术.软件学报,2007,18(2):246–258.

2 Tung A, Lu H, Han J, et al. Breaking the barrier of transactions: mining inter-transaction association rules. Proc. of the Knowledge Discovery and Data Ming. 1999.

3 秦亮曦,史忠植.时间序列跨事务关联分析研究.计算机工程与应用,2005,27(41):10–12,173.

4 Lu H, Han J, Feng L. Stock movement and n-dimensional inter-transaction association rules. Proc. of the SIGMOD Workshop on Research Issues on Data Ming and Knowledge Discovery. 1998.

5 Lu H, Feng L, Han J. Beyond intra-transaction association rules. ACM Trans. on Information Systems, 2000, 18(4): 423–454.

6 Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: Dunham M, Naughton J, Chen W eds. Proc. of 2000 ACM-SIGMOD Int'l Conf on Management of Data(SIGMOD'00). Dallas, TX, New York. ACM Press. 2000. 1–12.

7 Han JW, Kamber M. 范明,孟小峰,等译.数据挖掘:概念与技术.第 2 版.北京:机械工业出版社,2007.

8 范明,李川.在 FP-树中挖掘频繁模式而不生成条件 FP-树. 计算机研究与发展,2003,8(40):1216–1222.

9 李小兵,吴锦林,薛永生,翁伟.关联规则挖掘算法的改进与优化研究.厦门大学学报(自然科学版),2005,4:71–74.

10 宋余庆,朱玉全.基于 FP-Tree 的最大频繁项目集挖掘及更新算法.软件学报,2003,14(9):1216–1222.